STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT DOCUMENTATION PAGE

DRISI-2011 (REV 10/1998)

1. REPORT NUMBER	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER
CA-17-2985		

4. TITLE AND SUBTITLE	5. REPORT DATE
UTC-Enabling Demand Modeling from Privately Held Mobility Data	06/30/2017
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR	8. PERFORMING ORGANIZATION REPORT NO.
Alexey Pozdnukhov, Madeleine Sheehan, Mogeng Yin	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. WORK UNIT NUMBER
University of California at Berkeley Institute of Transportation Studies Berkeley, CA 94720	11. CONTRACT OR GRANT NUMBER 65A0529 TO 048
12. SPONSORING AGENCY AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
California Department of Transportation Division of Research, Innovation and System Information PO Box 94873, MS 83 Sacramento, CA 94273-0001	Final Report 5/1/16 - 4/30/17
	14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

This paper presents the design of the travel mode detection component within a generic architecture of processing individual mobility data. It approaches mode detection in two steps, each aiming at a particular objective. The first step develops a discriminative classifier that detects the mode of the observed trips or a sequence of modes in a multiple leg journey. It requires a considerable amount of ground truth data with known modes to be available for training. It also relies on a k-shortest path algorithm that generates plausible alternatives routes for the journey. The second step utilizes the discriminative recognition step of the observed mode in order to build a behaviorally grounded model that predicts the chosen mode within a set of available alternatives as a function of user characteristics and transportation.

17. KEYWORDS	18. DISTRIBUTION STATEMENT	
Travel Demand Models (TDMs), origin destination, Activity Based travel demand Models (ABMs), Discrete Choice Model (DCM), Long Short Term Memory (LSTM),		
19. SECURITY CLASSIFICATION (*of this report*)	20. NUMBER OF PAGES 10	21. COST OF REPORT CHARGED

Disclaimer Statement

This document is disseminated in the interest of information exchange. The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California or the Federal Highway Administration. This publication does not constitute a standard, specification or regulation. This report does not constitute an endorsement by the Department of any product described herein.

For individuals with sensory disabilities, this document is available in Braille, large print, audiocassette, or compact disk. To obtain a copy of this document in one of these alternate formats, please contact: the Division of Research, Innovation, and System Information, MS-83, California Department of Transportation, Division of Research, Innovation, and System Information, P.O. Box 942873, Sacramento, CA 94273-0001.

University of California
Center for Economic
Competitiveness in
Transportation

UCONNECT

Enabling Demand Modelling from Privately Held Mobility Data

Final Report

Alexey Pozdnukhov, UC Berkeley

Sponsored by

Caltrans

2017

Berkeley UC Irvine UCLA UCR UCSB CAL POLY POMONA

Tasks 3 and 4: context recognition and alternatives set generation for travel mode detection, DCM parameter calibration.

Madeleine Sheehan
UC Berkeley
m.sheehan@berkeley.edu

Mogeng Yin
UC Berkeley
mogengyin@berkeley.edu

Alexei Pozdnoukhov
UC Berkeley
alexeip@berkeley.edu

ABSTRACT

This papers presents the design of the travel mode detection component within a generic architecture of processing individual mobility data. It approaches mode detection in two steps, each aiming at a particular objective. The first step develops a discriminative classifier that detects the mode of the observed trips or a sequence of modes in a multiple leg journey. It requires a considerable amount of ground truth data with known modes to be available for training. It also relies on a k-shortest path algorithm that generates plausible alternatives routes for the journey. The second step utilizes the discriminative recognition step of the observed mode in order to build a behaviorally grounded model that predicts the chosen mode within a set of available alternatives as a function of user characteristics and transportation system variables. It is based on the discrete choice modelling paradigm and results in a set of parameters calibrated for distinct neighborhoods and/or segments of population. The overall framework therefore enables travel mode choice modeling and a consequent policy analysis and transportation planning scenario evaluation by leveraging privacy-sensitive individual mobility data possibly held in a secure private repository. It provides a set of algorithms to drastically reduce the latency and costs of obtaining a crucial information for models used in transportation planning practices. The performance and accuracy of the algorithms is evaluated experimentally within a large metropolitan region of the San Francisco Bay Area.

Keywords

UPDATE

1. INTRODUCTION

Travel Demand Models (TDMs) are an important tool for transportation planning. TDMs typically rely on travel surveys that are expensive, infrequent, and slow to reflect changes to the transportation system. Recent studies have proposed methods for generating travel demand model inputs from passively collected location data from mobile phones. Several methods have focused on extracting origin-destination (OD) matrices, and perform either dynamic traffic assignment or simulation to estimate the traffic volume on road network, which corresponds to traditional trip based travel demand models. A few have attempted to model individual agent activities and trips, which corresponds to more advanced Activity Based travel demand Models (ABMs). ABMs are based on the idea that travel is derived from people's desire to complete activities. The activity is the nuclear unit of such a model; the ABM will predict what activities a person wants to partake in, when and where the activities will occur, and how the person will travel to each activity. ABMs typically assign travel mode probabilistically, according to the outputs of a Discrete Choice Model (DCM). The DCM parameters are typically derived from household travel survey responses.

Previous studies using cellular data to inform ABMs have achieved a good understanding of the activity (trip purpose) patterns. However, the missing piece is travel mode and route inference. In this paper, we try to fill the gap by showing how passively collected big data sources can be used to infer the travel mode used to get from one activity to the next. Moreover, a DCM based on the inferred travel mode is trained so that such a model can be directly used for transportation planning.

In the Bay Area, on average, a phone accesses the network every 1.2 minutes. For long/non-trivial trips, a cell phone will typically create several CDR entries during travel. These records encode rich information about the spatial-temporal nature of the trip (i.e. travel speed, frequency of data records, proximity to road and transit infrastructure).

In this paper we show how to build an ABM that incorporates activity selection, location and time choice, and travel mode selection from passively collected cell phone data. We are limited (at present) by the availability of ground truth information on a traveler's selected travel mode. In this work we highlight two methods for dealing with the lack of ground truth and inferring the travel mode. Method one involves generating realistic cell records for simulated travel and building a classifier to determine the travel mode. We have at our disposal a well calibrated travel simulation tool for the 9 counties of the bay area. The simulator includes travel by by car, bus, train, subway, tram (light rail) and cable car. In this tool agents iteratively select travel alternatives until they find an optimal travel mode/route. We use a sequence-to-label Long Short Term Memory (LSTM)

LSTS 2016 August 14, 2016, San Francisco, CA, USA

DOI: 10.475/123_4

neural network ((add reference)) to learn the conditional probability of the cell phone user's mode of travel during a trip given the sequence of cell phone records created during travel. The input amounts to a sequence of timestamped latitude, longitude coordinates (one input vector for each cell record). The details of the method are discussed in [[Section Ref]]. The model allows for variable size input sequences, and automatically learns relevant properties of the travel modes (like location of infrastructure specific to one or more modes of travel, travel speeds, etc.).

The second method involves querying an external routing database to generate a list of travel alternatives. A naive Bayes approach is used to compare the observed cell records to the alternatives and determine the most likely mode. The method potentially enables better accuracy as the model explicitly compares the cell records to a set of possible alternatives - where the LSTM approach attempts to learn the selected travel mode with no information about the alternatives. However, the process of querying an external database to generate the set of travel alternatives for each trip in the dataset is, at present, infeasible. We demonstrate the effectiveness of this algorithm on a small sample of trips.

Finally we complete the ABM framework by training a DCM on the observed trips, inferred trip purpose, predicted travel modes and observable characteristics of a traveler. The demonstarted DCM is simplistic, but the model could certainly incorporate unique features such as the user's typical travel mode (inferred from previous trips in the dataset), the travel mode of previous trips in the same activity chain (if someone drove to the store it is likely that they will also drive home), and trip purpose (as determined by the semi-supervised IO-HMM). The final output of our work is a generative model for producing realistic activity chains for users (including activity time and location choice) and a discrete choice model to inform how the user will travel to the locale.

While the interpretation of CDRs is less accurate than the information from travel surveys, we benefit from the vast coverage of cell phone network; cell providers have access to a much larger sample size than household travel surveys typically do. Due to small sample size ABMs often fit one global model to all travelers and all trips (regardless of trip purpose) in a region. With the vast number of customers that the cell phone network services we have enough information to train local ABMs and are not bound to the assumption that ABM parameters are constant over the whole region. (Add note about when paired with IOHMM activity model, DCM framework allows for population segmentation based on attributes like travelers inferred home location or the travelers lifestyle (i.e. whether the traveler is a regular everyday home-work-home commuter, whether she typically partakes in secondary activities on the way home from work, or whether she frequently telecommutes)),

The remainder of this paper is organized as follows. Section 2 gives a literature review of related works on mode detection methods. Section 3 describes the discriminative mode recognition step, introducing a range of models and discussing their applicability and algorithmic constraints given available data. In Section 4, a discrete choice modelling framework is introduced. Section 5 presents an experimental evaluation of the framework across the range of performance metrics. Finally, in Section 6, we draw our conclusions and present the future work.

2. RELATED WORK

2.1 Travel Mode Detection from Passively Collected Data

While several studies have used GPS data to infer travel route and mode [7]. To date, there have been few studies that have used CDR traces to infer transportation route, and fewer still that use CDR data to infer transportation mode.

GPS locations are generally more accurate than CDR locations. However, GPS services do not give the same population coverage or the consistent temporal coverage that you get from CDRs. GPS locations are collected by a cellphone application provider while the application is enabled (if the user has enabled location based services). In other words, application service providers have access to GPS data only for the apps' user base and only when users are using the app and have enabled location services. Apps are generally not enabled at all times - meaning there are large gaps in coverage. Some studies have sought volunteers to enable GPS and be constantly monitored for transportation survey purposes, but generally this is for a very small sample. [liao2007,]

CDRs, on the other hand, are automatically collected for all of the carriers' customers. As mentioned above, a record is created anytime a phone places or receives a call, text, or accesses data. CDRs, therefore, offer broad population coverage - a single service provider typically provides service to 25-40% of the population in an area. CDRs also benefit from broad temporal coverage - on average, a phone creates a CDR record every ?? minutes.

2.1.1 Travel Mode Detection from GPS

At this point, travel mode detection from GPS data has become quite mature. The task has received more attention than classification from raw CDR data because it has higher spatial and temporal resolution. Most studies using GPS are experimental and small scale, and have the ground truth travel mode labels to train a supervised classifier.

Source of ground truth The automated transportation and emissions calculator app called E-missions tracks 44 users for 3 months and obtains ground truth travel labels by asking the user to confirm their travel mode once the trip has ended. [9]. Table 1 of Shankari et al. gives a summary of other GPS based travel mode detection experiments. The number of praticipants in the experiments ranges from 5-135. [9].

Features Several experiments use pure data from GPS [1] [13], [14]. Accelerometer data provide a highly discriminative feature of motion dynamics with distinct signatures for each of the modes. [8] and improves the accuracy of travel mode detection by 17% over using GPS alone. Others incorportates GIS information in addition to GPS and accelerometer data [10]. These experiments take advantage of features like average speed, average acceleration, distance from bus routes, and even use realtime transit feeds to determine average candidate bus closeness.

Models GPS based mode detection algorithms take advantage of many models including Neural networks [4], Decision trees [14], SVM [14], graphical model [6]

2.1.2 Mode Detection from CDRs

Holleczek et al. extracted trips from CDR data in Singa-

pore. They then quantified the travel mode split (driving vs. public transport) by comparing the CDR generated OD matrix to the public transport data generated from smart card access. (Use this study to identify service gaps)

Doyle et al. classified the travel mode of users travelling between Dublin and Corkbetween in the Republic of Ireland using CDR data [3]. They measured the likelihood of an observed trajectory being a road trip or a rail trip based on the proportion of locating events that occur at cells that represent the route of interest. However, the limitation of their paper is that 1) they only classified the travel mode between car and rail, 2) they only considered one pair of origin and destination, and 3) they only used spatial features of the CDR data.

Wang et al. grouped trips by their origin and destinations (by 500*500 cells). They clustered the trips in each group by the travel time using K-Means algorithm [11]. The faster cluster of trips are assigned as driving trips, the slower cluster is assigned public transit. They validated the cluster mean travel mode with the Google map travel time for each origin and destination group. However, their method did not use any way points during the trip thus any spatial features. The selection of 2 clusters per group is not justified and there is no direct validation presented.

Yoo et al. compared the estimation of travel time using GPS and cellular data. Travel mode identification was used as a pre-processing step prior to map matching. Their travel model inferences are purely rule-based, i.e., pedestrian walks below 5 kph speed. Bus or train would stops at stops or stations, etc.[2]

Leontiadis et. al devised an algorithm to infer the mobility path between activity locations based on cellular network topology and GIS information. They used A* algorithm searching from the GIS road networks, with the weight of the road biased to the high-probability paths based on the observed way points. Their result also show a median accuracy of 70m compared with ground truth GPS trajectories. They also show that mobility path accuracy improves with its length and speed [5].

Route detection from CDR and Handover Data. Tettemani et al. use cell handover (HO) data to predict the probable route of a traveler. Handovers transfer an ongoing call/data-session from one tower to the next if a phone is moving while a call is in session. If a call or data session lasts for the entirety of the trip, the HO data indicates every towers accessed along a route. Tettemani et al. predict the route from a distance measure from a possible route (path) to the centroid of the cell zones for cells accessed while en route. However this work is limited, because, as mentioned, handover data relies on the phone accessing the network for the entirety of the trip.

Wu et al. propose a method to estimate route flow (the number of vehicles using each route to travel from an origin to a destination) using a combination of link flows from traditional traffic sensors and data from the cell network. Each possible route from an origin to destination is associated with a cellpath - a sequence of towers accessed while on a particular route. While the paper doesn't say so explicitly, their method relies on handover like data. They assume that there is only one cell-path associated with each route.

CDR data does not require a constant session. [[TODO: add details]]

Benefits of my method: - relies on cell phone data - covers greater user base than GPS data -allows for population segmentation in ways that explain heterogeneity in travel mode choice. - CDR data rather than HO data (does not require a session to be active for the entirety of a trip) - Link flows from Traffic sensors and Public Transport ridership info can be used to validate findings, but is not inherently needed for route classification or for calculating mode split.

2.2 Mode Detection and Map Matching

Map matching, particularly related to travel on networks, is a problem of associating a set of observed coordinates of a moving object with a sequence of links that this travel takes place on, either offline [] or online []. Efficiency and accuracy of map matching algorithms are at the cornerstone of multiple data processing systems producing travel related information from location data, and GPS probe data in particular[]. With decreasing spatial localization accuracy, the nature of the problem changes to route flow inference[]. Map matching is closely related to mode detection as map matching algorithms, whether offline or online, can be used to infer which link of a multi-modal network the travel takes place on, and, combined with speed information, inform mode detection. The impact of map matching approaches is profound when multiple modes take geographically distinct routes [3], and diminishes when multiple modes share a spatial corridor within the localization accuracy of the sensing technology. The presented work utilizes elements of map matching algorithms of [].

3. METHODOLOGY

In this work we have two main objectives - the first is to build a discriminative model to automatically detect a person's mode of travel during a trip. The second objective is to fit a discrete choice model to the observed travel mode. The parameters of the discrete choice model describe and explain how travelers choose the travel mode from a discrete set of travel mode alternatives and can be used to predict what travel modes a traveler will take on future trips.

For the discriminative portion we suffer from a lack of ground truth information on what travel mode is actually taken for a given trip. We develop two methods for classifying the travel mode - each has a different approach for dealing with the lack of ground truth. In the first, non-parametric model, we generate realistic cell records for simulated travel and use a sequence to label LSTM to convert the sequence of observed cell records into a travel mode label. In the second, we rely on an external routing database to provide information on the availabl travel alternatives for a given trip, and use a naive Bayes (parametric) approach to determine the most likely alternatives from the set.

Finally we train a discrete choice model using the predicted travel mode, properties of each of the travel alternatives, and characteristics of the travelers. The details of each step are outlined below.

3.1 Discriminative model 1: Sequence to label LSTM neural net

3.1.1 Realistic simulation of CDRs from observed trips

The LSTM training relies on a regional micro-simulation to provide the location and activity of each agent throughout a day, including realistic travel by car, bus, train, subway, tram, light rail, and cable car. We simulate cell records

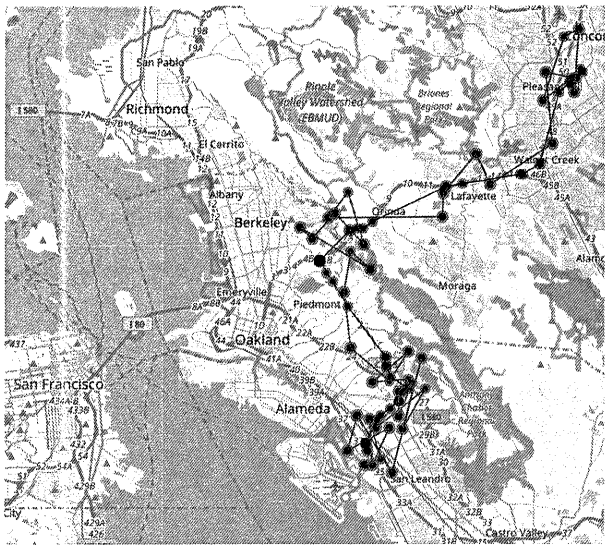

Figure 1: Simulated cell records for a car trip: The black dots represent the location of the agent at the time of the simulated record is shown in black. The corresponding simulated cell records are shown in blue

along the agents path of travel. The simulated cell records amount to a (timestamp, latitude,longitude) tuple with location noise.

The average rate at which cell records are created during travel is about 50 records per hour. The reported location accuracy is about 1km. Therefore we simulate cell records along the travel path with 1 km of Gaussian noise. Figure 1 shows an example of simulated cell records for a driving trip from the micro-simulation. We use this procedure to simulate CDRs for 48,000 bay area trips.

3.1.2 *Training the LSTM Neural Network*

After simulating CDRs along the agents travel paths we train an LSTM sequence to label neural network to learns the conditional probability of each travel mode given the sequence of latitudes, longitudes, and timestamps observed during the trip.

While this model is trained on simulated CDR data, the same could certainly be trained on sequences of coordinates and timestamps observed from GPS or other sources assuming that the travel mode labels are known. The LSTM automatically learns relevant features like location of travel-mode specific infrastructure, possibly the travel speeds or prolonged stops of buses.

The LSTM inputs are normalized: the location coordinates and timestamps are scaled to be between 0 and 1. These inputs are fed into an encoder followed by 2 hidden layers each with 128 nodes. The output of the hiddent nodes are fed to a 6×1 output layer. We perform a softmax at the output layer to determine the probability of each travel mode: car, bus, train, subway, tram, light rail, and cable car.

3.2 Model 2: Naive Bayes approach

3.2.1 *Stay point detection from CDRs*

The goal of stay location recognition is to turn CDR logs

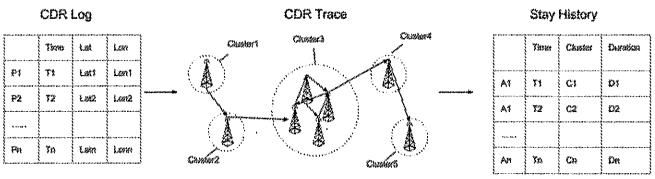

Figure 2: Call Detail Records (CDR) data collection

into a list of sequential stay locations with start time and duration for each user, as illustrated in Figure 2. Each record of raw CDR logs contains the timestamp and the approximate latitude and longitude of events recorded by the data provider. This is a CDR-specific step that requires fine-tuning of several threshold parameters. The details of this algorithm are outlined in [12]. After we have identified the stay locations, a trip is defined as travel between two consecutive stay points.

3.2.2 *Building alternative set*

For each trip we query a multi-modal routing database (such as Google maps or Open Street Maps) to obtain a set of driving, transit, biking and walking (where appropriate) route alternatives. For each alternative we retain attributes of the trip:

- travel mode
- total travel time
- route geometry
- expected travel time
- specifically for transit:
  - number of transfers
  - type of transit
  - walking access/egress distance
  - agency names
  - route name/number

These features are used both in the parametric discriminative travel model and several are also relevant for fitting a DCM.

3.2.3 *Computing likelihood of alternatives*

We model the likelihood of each alternative using a naive Bayes classifier. We want to compute the probability of each travel alternative, y_k, in the alternative set given the sequence of CDRs that we observe: $x_1 \ldots x_n$:

$p(y_k \mid x_1, \ldots, x_n)$

We take advantage of Bayes rule to decompose $p(y_k \mid \mathbf{x})$ as follows:

$$p(y_k \mid \mathbf{x}) = \frac{p(y_k)\ p(\mathbf{x} \mid y_k)}{p(\mathbf{x})}$$

The denominator - the probability of observing the sequence of cell phone records that we observe, is constant for all travel modes in the set, so

$$p(C_k \mid \mathbf{x}) \propto p(C_k)\ p(\mathbf{x} \mid C_k)$$

Here we treat the input sequence as a measure of true location with location nosie. The probability of observing a record, x_i at a distance $d_{i,k}$ from route k is:

$$\Pr(x_i \mid y_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(d_{i,k})^2}{2\sigma_d^2}}$$

3.3 Discrete Choice Model

With travel mode inferred we can formulate a parametric discrete choice model to learn traveler's mode preferences and determine how traveler's trade off various attributes of the travel alternatives.

In a multinomial DCM, the utility, U, that person n obtains from choosing alternative i, depends on attributes of each travel alternative in the choice set:

$$U_{ni} = \beta_j z_{in} + \epsilon_{ni}$$

where z_{ni} represents a vector of observed variables of trip i and traveler n, β is a vector of the corresponding coefficients that are interacted with each of the observable variable, and $\epsilon_n i$ represents the unobservable factors that contribute to travel mode choice.

The probability of traveler n selecting travel alternative i is the probability that the utility of alternative i is greater than the utility of all other alternatives. If there are J total alternatives, the probability that mode i is given by the following:

$$P_{ni} = \frac{exp(\beta z_{ni})}{\sum_{j=1}^{J} exp(\beta z_{nj})})$$

We consider a representative sample of Bay Area trips. The coefficient β are found by maximum likelihood estimation.

4. EXPERIMENTAL EVALUATION

There is no ground truth to directly quantify the accuracy of individual activity assignments of our proposed discriminative models for real cell records. However, the travel micro-simulation tool allows us to evaluate the discriminative models' ability to recover the travel mode for simulated travel and realistic cell records along the travel path.

In order to assess the models performance on actual cell records, we use our methods to infer travel mode split in the region and on key commuting trips. We compare our model with aggregated statistics from surveys. Finally we evaluate the parameters of the DCM to see if the parameters compare well to those typically seen in travel demand models.

4.1 Discriminative model

4.1.1 LSTM for mode detection

The non-parametric LSTM model is trained on simulated cell records. We use 80% of the micro-simulation generated trips to train the model and 20% as a test set. The dataset consisted of 48,000 trips. The distibution of travel mode on these trips is shown in Figure ??. As the data-set is highly imbalanced, we use an imbalanced cross entropy loss function that penalizes missed trips of the transit trips more heavily than it penalizes a missed drive trip.

After every 50 training batches we compute the accuracy on test set. Figure 4 shows us the per travel mode recall. Recall is a machine-learnin and statistical term that denotes the fraction of samples in a given class that are correctly

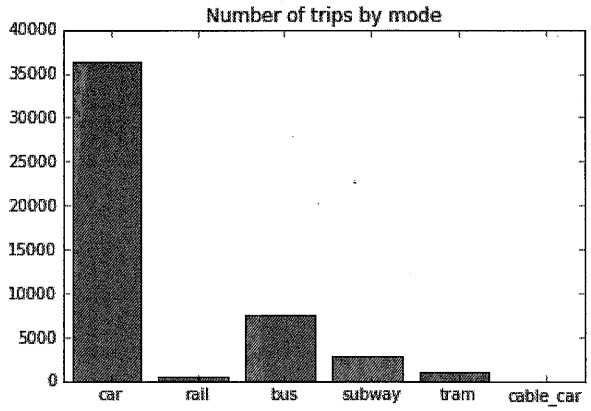

Figure 3: LSTM dataset - number of trips by travel mode generated from travel micro-simulation tool.

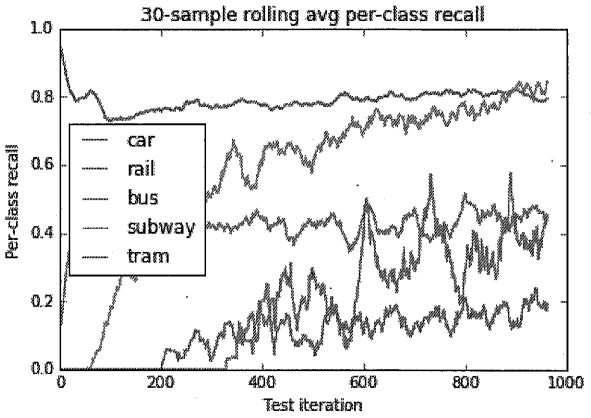

Figure 4: LSTM per class recall on the test set. Recall is evaluated after every 50 training batches

identified as belonging to that class. The recall for drive trips, for example, tell us what percentage of the drive trips were actually labeled as drive trips. The precision, on the other hand, represents the fraction of samples that are correctly labeled as belonging to a class over the total number of samples labeled as belonging to that class. The per-class precision is shown in Figure 5

As seen in Figure 4, the recall for car trips begins near 1.0. In early training the neural net predicts that all trips belong to the most prominent travel mode in the training set - in this case the car mode. As the training continues, however, the precision for the other modes improves. In particular the model achieves a very high subway recall by the end of training.

The relatively low precision of all non-car travel modes indicates that the model over-predicts transit trips.

Figure 6 shows the confusion matrix at the end of training highlighting which travel modes are misclassified. The model still struggles to differentiate between bus and drive trips and also often confuses tram trips with bus or drive. These make sense as these travel modes often share infrastructure.The rail an subway often have dedicated track that is often spatially separated from the road network.

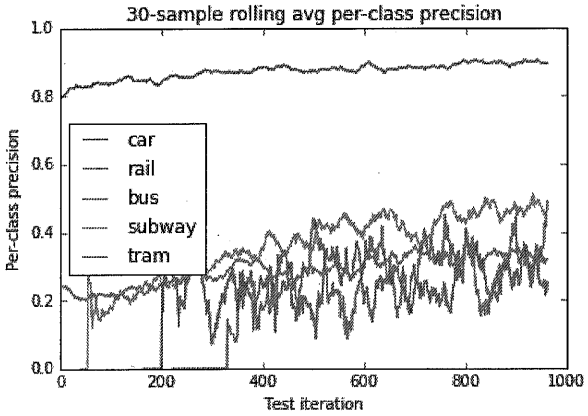

Figure 5: LSTM per class precision on the test set. Precision is evaluated after every 50 training batches

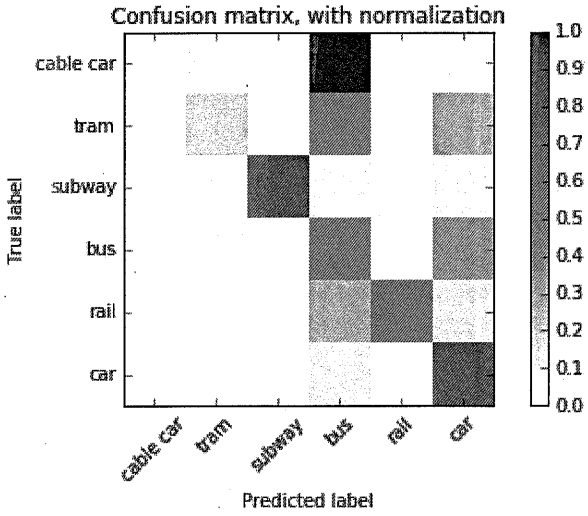

Figure 6: After training, we obtain the following confusion matrix between travel modes in the test set.

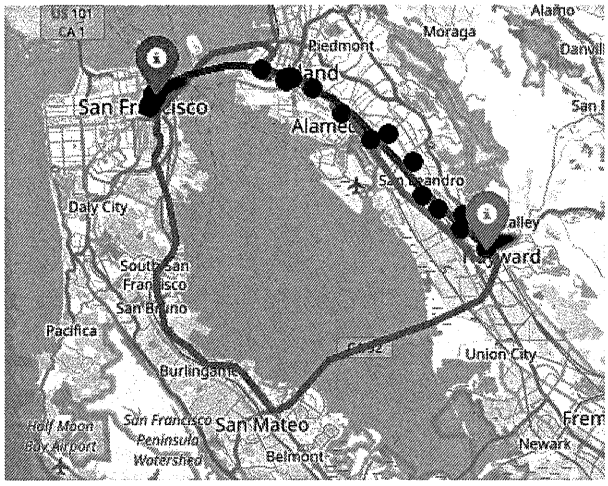

Figure 7

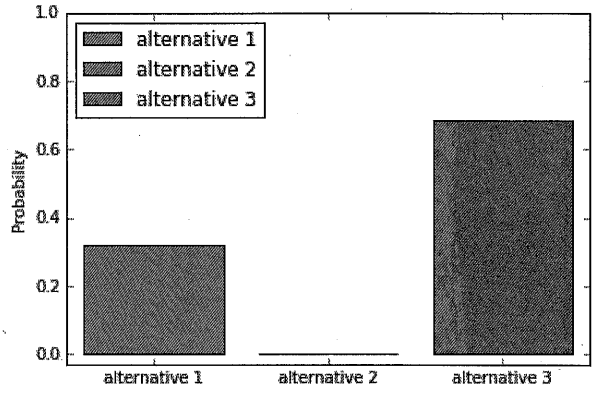

Figure 8

4.2 Naive Bayes parametric discriminative model

Due to privacy protection policies enforced by the CDR data provider, there is no direct ground truth to quantify the accuracy of individual activity assignments by our proposed model. On routes where there is spatial distinction between the alternatives the classifier naive bayes approach works well. In dense urban areas where there is less separation between alternatives, the method may not work as well. Figure 7 shows a set of alternatives for a given route and Figure 8 gives the probability of each alternative when we use σ_d of 1 km.

4.3 Discrete Choice Model Results

Either of the above methods (or other discriminative methods) can be used to infer the travel mode. Using the most likely travel mode from the naive bayes classifier, we fit a simple DCM to the observed travel modes. We infer the travelers home location according to the methods outlined in [12] and use the median income of the traveler's home census tract as a proxy for the traveler's income. We query an in-house routing service that provides travel times and costs for a set of possible travel alternatives between the observed origin and destination zones. Eq. (??) shows a DCM specification that accounts for the time and cost of travel, a traveler's anticipated income.

Table 1: Discrete choice model parameters for travel mode

Variable	Coeff	Std. error	Z	P > |Z|
β_{drive}	-1.048	0.525	-1.996	0.046
β_{income}	0.0156	0.006	2.557	0.011
β_{TT}	-1.9495	0.413	-4.72	0.000
β_{TC}	-0.0653	0.061	-1.071	0.284

$$\begin{aligned} V_{drive} =& \beta_{drive} + \beta_{income} * Income \\ &+ \beta_{TT} * TravelTime_{drive} \\ &+ \beta_{TC} * TravelCost_{drive} \\ V_{public\_transit} =& \beta_{TT} * TravelTime_{public\_transit} \\ &+ \beta_{TC} * TravelCost_{public\_transit} \end{aligned}$$

this model resulted in parameters listed in Table 1.

5. ACKNOWLEDGMENTS

This work was partially funded by a gift from AT&T. Support from the State of California Department of Transportation (CalTrans) through UCCONNECT faculty research grant program, agreement 65A0529, is also acknowledged.

6. REFERENCES

[1] R. Brunauer, M. Hufnagl, K. Rehrl, and A. Wagner. Motion pattern analysis enabling accurate travel mode detection from gps data only. pages 404–411, 2013.

[2] Y. Byeong-Seok, S. Samsung, K. Seung-Pil, and C.-H. Park. Travel time estimation using mobile data. In *Proceedings of the Eastern Asia Society for transportation studies*, volume 5, pages 1533–1547. Citeseer, 2005.

[3] J. Doyle, P. Hung, D. Kelly, S. McLoone, and R. Farrell. Utilising mobile phone billing records for travel mode discovery. 2011.

[4] P. Gonzalez, J. Weinstein, S. Barbeau, M. Labrador, P. Winters, N. L. Georggi, and R. Perez. Automating mode detection using neural networks and assisted gps data collected using gps-enabled mobile phones. In *15th World congress on intelligent transportation systems*, page 2008, 2008.

[5] I. Leontiadis, A. Lima, H. Kwak, R. Stanojevic, D. Wetherall, and K. Papagiannaki. From cells to streets: Estimating mobile paths with cellular-side data. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, pages 121–132. ACM, 2014.

[6] L. Liao, D. Fox, and H. Kautz. Blocation-based activity recognition using relational markov networks,[in proc. 19th int. In *Joint Conf. Artif. Intell., Edinburgh, Scotland*, pages 773–778, 2005.

[7] L. Liao, D. Fox, and H. Kautz. Extracting places and activities from gps traces using hierarchical conditional random fields. *The International Journal of Robotics Research*, 26(1):119–134, 2007.

[8] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks (TOSN)*, 6(2):13, 2010.

[9] K. Shankari, M. Yin, S. Shanmugam, D. E. Culler, and R. H. Katz. E-mission: Automated transportation emission calculation using smart phones. *Electrical Engineering and Computer Sciences University of California at Berkeley*, 1, 2014.

[10] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu. Transportation mode detection using mobile phones and gis information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 54–63. ACM, 2011.

[11] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 318–323. IEEE, 2010.

[12] M. Yin, M. Sheehan, S. Feygin, J.-F. Paiement, and A. Pozdnoukhov. A generative model of urban activities from cellular data.

[13] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 312–321. ACM, 2008.

[14] Y. Zheng, L. Liu, L. Wang, and X. Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web*, pages 247–256. ACM, 2008.